A rapid data acquisition pipeline for visualizing the 2009 A/H1N1 influenza pandemic

DONOVAN H. PARKS, NORMAN J. MACDONALD, AND ROBERT G. BEIKO
Faculty of Computer Science, Dalhousie University

Key words: data mining, GenGIS, H1N1pdm, influenza A, information visualization

1. INTRODUCTION

The 2009 A/H1N1 influenza pandemic (H1N1pdm) has spread to nearly all parts of the world and continues to be of prime concern to public-health officials as we enter our traditional flu season. Advances in DNA sequencing technology have allowed thousands of H1N1pdm isolates to be sequenced with new isolates being added daily to the NCBI Influenza Virus Resource database. This extensive sequencing effort has allowed researchers to estimate when the H1N1pdm virus was first transmitted to humans (Rambaut and Holmes, 2009), to reconstruct the initial spatiotemporal dynamics of the pandemic (Parks *et al.*, 2009b; Lemey *et al.*, 2009; Jombart *et al.*, 2009), and to assess the transmissibility and severity of this new strain (Fraser *et al.*, 2009). Keeping such analyses up-to-date is essential for informing public health policy. Here we present an automated pipeline for retrieving and parsing H1N1pdm isolates along with a geospatial visualization package for interactively exploring the spatiotemporal progression of new mutations and viral strains. We demonstrate the utility of these research tools by examining (1) the geospatial evolutionary history of the H1N1pdm and (2) the spatiotemporal dynamics of a pair of amino acid changes (i.e., polymorphisms) in the neuraminidase (NA) protein of influenza A that are implicated in antibody recognition. These tools allow continual monitoring of the evolution of influenza A which is essential for the early identification of new outbreaks and for moving towards preventive health strategies such as early containment (Cooper *et al.*, 2006; Ferguson *et al.*, 2006).

2. METHODS

Data Acquisition. We have implemented a custom Python pipeline that can automatically acquire and parse new H1N1pdm records from the Influenza Virus Resource at NCBI. H1N1pdm records contain sequence information for an isolate along with an EpiFlu annotation block specifying metadata such as the location and time of collection. Our pipeline uses MUSCLE version 3.7 (Edgar, 2004) to add new sequences to an existing H1N1pdm multiple sequence alignment and RAxML version 7.0.4 (Stamatakis, 2006) to infer a maximum-likelihood tree indicating the evolutionary relationship between isolates. As the quality and completeness of annotations varies extensively between sequencing labs, we infer isolate locations and collection dates from context (e.g., isolate name) wherever possible. Location names are geo-tagged with latitude/longitude coordinates in a semi-automated manner by querying the GeoNames server (http://www.geonames.org/). The output of this data acquisition pipeline is a phylogenetic tree containing leaf nodes with known geographic locations and a comma separated values (CSV) file containing the metadata known about each isolate.

Data Visualization. We have developed a geographic information system, named GenGIS, for visualizing and analyzing genetic data where identifying spatial or temporal patterns can help elucidate underlying processes such as the transmission dynamics of viral pandemics (Parks *et al.*, 2009a). As a free and open source application, GenGIS makes extensive use of other open source software libraries, including Python which provides GenGIS with a powerful scripting language, RPy which contains extensive support for performing statistical tests and analyses, and GDAL which allows map data in nearly any format to be imported and manipulated. GenGIS has been implemented in C++ in order to support the computationally efficient handling of large datasets and uses

Correspondence should be addressed to DHP (parks@cs.dal.ca).

OpenGL for 3D rendering. In addition to the visualizations and data analysis options implemented directly in GenGIS, users can extend its functionality using the application programming interface (API) exposed to the built-in Python interpreter.

For this analysis, we have implemented a Python script that uses the GenGIS API to generate site-by-site pie charts indicating the relative abundance of different polymorphic sites at varying times. This allows the spatiotemporal dynamics of a polymorphism to be studied. Emergent polymorphisms which may result in increased severity or virulence can be tracked on a daily basis using the CSV file produced by our data acquisition pipeline. We also investigate the evolutionary relationship of sequences collected from different geographic locations using both a traditional three-dimensional 'geophylogeny' (Janies *et al.*, 2007; Kidd and Ritchie, 2006) and a novel two-dimensional tree representation. This two-dimensional representation allows users to define an explicit geographic axis and quantitatively assess how well their data fits this axis (Parks and Beiko, 2009). These tree visualizations provide insight into the transmission dynamics of the H1N1pdm.

3. RESULTS

In this paper, we demonstrated how combining an interactive visualization environment with an automated data acquisition pipeline allows the geospatial evolution of the H1N1pdm and the spatiotemporal dynamics of polymorphic sites to be efficiently monitored. Visualizing the relative abundance of different polymorphic sites over time and space allowed us to identify a pair of polymorphic sites in the NA protein implicated in antibody recognition which exhibit strong geographic structuring. Notably, the putative ancestral strain for these polymorphic sites is almost exclusively isolated to the Asian countries of China, South Korea, and Japan. The sole exception is a single isolate from Italy. Should this strain or a descendant from it prove to be of high severity or virulence, it may dictate minimizing travel to and from these countries.

Widespread travel coupled with delayed symptomatology has permitted the H1N1pdm virus to rapidly spread around the global. Visualizing 2D and 3D geophylogeny within GenGIS suggests that the virus has been spreading extensively before it was first identified in April 2009 and has been independently introduce to most geographical regions multiple times. These results indicate that more extensive sampling is required if we hope to implement health strategies such as early containment. However, continual monitoring after a pandemic has occurred is still desirable to global health-professionals in case a new viral strain requires containment protocols to be implemented.

4. CONCLUSIONS

Early containment is recognized as an essential preventive health strategy for minimizing the impact of infectious diseases with pandemic potential. The high economic cost coupled with the loss of personal freedom resulting from containment strategies dictates that such decisions be based on robust, up-to-date scientific evidence. Our data analysis pipeline and visualization environment allow health officials to monitor new mutations within the influenza virus which may result in new pandemic strains. We plan to extend our framework in the future to incorporate monitoring of other infectious diseases such as HIV-1 and hepatitis C. Automated tools for monitoring infectious diseases are becoming increasingly important as sequencing technology permits levels of sampling that exceed the capacity of manual monitoring efforts.

REFERENCES

Cooper, B. S., Pitman, R. J., Edmunds, W. J., and Gay, N. J. (2006). Delaying the international spread of pandemic influenza. *PLoS Med*, **3**(6), e212.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**(5), 1792–1797.

Ferguson, N. M., Cummings, D. A. T., Fraser, C., Cajka, J. C., Cooley, P. C., and Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature*, **442**(7101), 448–452.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Kerkhove, M. D. V., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., Alpuche-Aranda, C. M., Chapela,

I. B., Zavala, E. P., Guevara, D. M. E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., and Collaboration, W. H. O. R. P. A. (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*, **324**(5934), 1557–1561.

Janies, D., Hill, A. W., Guralnick, R., Habib, F., Waltari, E., and Wheeler, W. C. (2007). Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). *Syst Biol*, **56**(2), 321–329.

Jombart, T., Eggo, R. M., and Dobb, P. (2009). Spatiotemporal dynamics in the early stages of the 2009 A/H1N1 influenza pandemic. *PLoS Currents: Influenza*, **RRN1026**.

Kidd, D. M. and Ritchie, M. G. (2006). Phylogeographic information systems; putting the geography into phylogeography. *J. Biogeography*, **33**, 1851–1865.

Lemey, P., Suchard, M., and Rambaut, A. (2009). Reconstructing the initial global spread of a human influenza pandemic. *PLoS Currents: Influenza*, **RRN1031**.

Parks, D. H. and Beiko, R. G. (2009). Quantitative visualizations of hierarchically organized data in a geographic context. In *Geoinformatics*, Fairfax, VA.

Parks, D. H., Porter, M., Churcher, S., Wang, S., Blouin, C., Whalley, J., Brooks, S., and Beiko, R. G. (2009a). GenGIS: A geospatial information system for genomic data. *Genome Res*.

Parks, D. H., MacDonald, N. J., and Beiko, R. G. (2009b). Tracking the evolution and geographic spread of influenza A. *PLoS Currents: Influenza*, **RRN1014**.

Rambaut, A. and Holmes, E. (2009). The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Currents: Influenza*, **RRN1003**.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**(21), 2688–2690.